Discordance between different bioinformatic methods for identifying resistance genes from short-read genomic data, with a focus on *Escherichia coli*

1.1 Author names

Timothy J Davies [a, b], Jeremy Swan [a, b], Anna E Sheppard [a, b], Hayleah Pickford [a, b], Samuel Lipworth [a, b], Manal AbuOun [c], Matthew Ellington [b, e], Philip W Fowler [a], Susan Hopkins [b, e], Katie L Hopkins [b, f], Derrick W Crook [a, b, d], Tim EA Peto [a, b, d], Muna F Anjum [c], A Sarah Walker [a, b](*), Nicole Stoesser [a, b, d] (*).

* contribution considered equal

1.2 Affiliation

a) *Nuffield Department of Medicine, Oxford University, Oxford, United Kingdom*
b) *National Institute for Health Research (NIHR) Health Protection Research Unit on Healthcare Associated Infections and Antimicrobial Resistance at University of Oxford, UK*
c) *Bacteriology, Animal and Plant Health Agency, Surrey UK*
d) *Oxford University Hospitals NHS Foundation Trust, Oxford, UK*
e) *Antimicrobial Resistance and Healthcare Associated Infections (AMRHAI) Division, UK Health Security Agency, London, UK*
f) *HCAI, Fungal, AMR, AMU and Sepsis Division, UK Health Security Agency, London, UK*

1.3 Corresponding author

Dr Timothy Davies, timothy.davies@ndm.ox.ac.uk
Alternate corresponding author:
Dr Nicole Stoesser, nicole.stoesser@ndm.ox.ac.uk

1.4 Keyword

Antimicrobial resistance genotyping, genomics, Escherichia coli, resistance prediction

1.5 Repositories:

Sequencing data are available at the following NCBI BioProject accession number: PRJNA540750.

2. Abstract

Several bioinformatics genotyping algorithms are now commonly used to characterise antimicrobial resistance (AMR) gene profiles in whole genome sequencing (WGS) data, with a view to understanding AMR epidemiology and developing resistance prediction workflows using WGS in clinical settings. Accurately evaluating AMR in Enterobacterales, particularly *Escherichia coli*, is of major importance, because this is a common pathogen. However, robust comparisons of different genotyping approaches on relevant simulated and large real-life WGS datasets are lacking. Here, we used both simulated datasets and a large set of real

E. coli WGS data (n=1818 isolates) to systematically investigate genotyping methods in greater detail.

Simulated constructs and real sequences were processed using four different bioinformatic programs (ABRicate, ARIBA, KmerResistance, and SRST2, run with the ResFinder database) and their outputs compared. For simulations tests where 3,092 AMR gene variants were inserted into random sequence constructs, KmerResistance was correct for all 3,092 simulations, ABRicate for 3,082 (99.7%), ARIBA for 2,927 (94.7%) and SRST2 for 2,120 (68.6%). For simulations tests where two closely related gene variants were inserted into random sequence constructs, ABRicate identified the correct alleles in 11,382/46,279 (25%) of simulations, ARIBA in 2494/46,279 (5%), SRST in 2539/46,279 (5%) and KmerResistance in 38,826/46,279 (84%). In real data, across all methods, 1392/1818 (76%) isolates had discrepant allele calls for at least one gene.

Our evaluations revealed poor performance in scenarios that would be expected to be challenging (e.g. identification of AMR genes at <10x coverage, discriminating between closely related AMR gene sequences), but also identified systematic sequence classification (i.e. naming) errors even in straightforward circumstances, which contributed to 1081/3092 (35%) errors in our most simple simulations and at least 2530/4321 (59%) discrepancies in real data. Further, many of the remaining discrepancies were likely “artefactual” with reporting cut-off differences accounting for at least 1430/4321 (33%) discrepants. Comparing outputs generated by running multiple algorithms on the same dataset can help identify and resolve these artefacts, but ideally new and more robust genotyping algorithms are needed.

3. Impact statement

Whole-genome sequencing is widely used for studying the epidemiology of antimicrobial resistance (AMR) genes in bacteria; however, there is some concern that outputs are highly dependent on the bioinformatics methods used. This work evaluates these concerns in detail by comparing four different, commonly used AMR gene typing methods using large simulated and real datasets. The results highlight performance issues for most methods in at least one of several simulated and real-life scenarios. However most discrepancies between methods were due to differential labelling of the same sequences related to the assumptions made regarding the underlying structure of the reference resistance gene database (i.e. that resistance genes can be easily classified in well-defined groups). This study represents a major advance in quantifying and evaluating the nature of discrepancies between outputs of different AMR typing algorithms, with relevance for historic and future work using these algorithms. Some of the discrepancies can be resolved by choosing methods with fewer assumptions about the reference AMR gene database and manually resolving outputs generated using multiple programs. However, ideally new and better methods are needed.

4. Introduction

Whole genome sequencing (WGS) has become a major tool for characterising the epidemiology of bacterial antimicrobial resistance (AMR) genes, representing a potentially highly discriminatory, non-targeted approach with significant advantages over other more targeted molecular techniques(1). In addition, WGS-based antibiotic susceptibility prediction has been successfully implemented as part of diagnostic and treatment workflows for *Mycobacterium tuberculosis*(2). Accurate WGS-based profiling of complete AMR gene content and prediction of susceptibility phenotypes would represent an attractive option for other commonly encountered clinical bacterial pathogens, such as Enterobacterales, including *Escherichia coli*.

Several key components are required for WGS-based AMR genotyping and predictions of susceptibility phenotype, including a robust AMR gene reference catalogue linking each genetic mechanism/sequence with a given phenotype, and accurate AMR gene identification and classification algorithms. Several catalogues and bioinformatics algorithms are now available(3-9), but only limited comparative evaluation of their outputs has been undertaken. The genetic mechanisms underpinning AMR in Enterobacterales and some other bacteria (e.g. *Pseudomonas aeruginosa*) are much more complex than those in *M. tuberculosis*, and whilst some studies suggest that WGS-based genotyping holds promise for AMR gene characterisation and the prediction of antimicrobial susceptibility for several different Enterobacterales species(10-12), the limited reproducibility and reliability of such methods in a blinded, head-to-head analysis across nine bioinformatics teams has been recently highlighted(13). However, this study was small (n=10 sequencing datasets, n=7 isolates), encountered a limited set of typing discrepancies, and used highly selected samples, meaning the impact of these issues on larger, real-world datasets remains unclear.

We therefore used simulations and three large, independent and diverse *E. coli* sequencing datasets to investigate the robustness and reproducibility of four widely-used WGS-based AMR genotyping methods (ABRicate, ARIBA, KmerResistance, and SRST2) at scale, investigating any encountered discrepancies.

5. Methods

AMR gene identification methods

We evaluated the impact of different bioinformatics tools using the same AMR gene catalogue, namely the ResFinder database (v.29/10/2019). At the time the study was designed (March 2018), to be included bioinformatics tools had to: (i) have publicly available code, (ii) run on local computing architecture without major modification, (iii) accept different AMR gene databases to ensure broad and long-term typing usability, and (iv) have a command line interface that could enable batch processing of large numbers of samples (**Table S1**).

We identified four publicly available bioinformatic tools that met these criteria and used distinct AMR gene identification approaches: ABRicate(14) (which searches for AMR genes in assemblies using BLASTn), SRST2(7) (which maps reads directly onto the formatted AMR gene database using Bowtie 2), ARIBA(6) (which combines

these two approaches, first mapping reads to the AMR gene database using minimap, and then creating local assemblies of the mapped reads using Fermi-lite) and KmerResistance(8) (which analyses shared k-mers between the query sequences and reference sequences in the AMR gene database) (**Fig.S1**). To mimic broad usability, each program was run using default parameters. For ABRicate, assemblies were first produced using SPAdes(15) run with default parameters.

Simulated data: single and multiple allele identification, and low coverage scenarios

Prior to evaluating real data, we considered the accuracy of each method in identifying known AMR gene alleles "inserted" into simulated flanking sequence constructs. For this, each AMR gene variant in the ResFinder database (n=3,092) was flanked by 1kb of random sequence (using Numpy v1.16.4(16) and combined using BioPython(17) v1.74) and reads simulated at 40x coverage using ART (details and rationale in Supplementary Methods, **Fig.1, S2**). Other ART parameters were: error profile="HISEQ2500", mean DNA fragment length (standard deviation)=480bp (150bp), and read length=151bp. Each bioinformatic method was then tested to see if it could correctly identify the AMR gene variant, using default parameters.

We also considered two *a priori* scenarios that are thought to affect AMR genotyping(18), namely a *multiple allele* scenario in which multiple closely genetically related alleles (see below) of a given AMR gene were present, and a *low quality* scenario reflected by low sequencing coverage. For the *multiple allele* scenario we excluded target AMR gene variants that were incorrectly identified individually by any method (see Results), and then calculated pairwise nucleotide similarity between all remaining AMR gene variants. To do this, each remaining AMR gene variant was split into 31-mers, which were then compared with 31-mer sets from every other non-excluded AMR gene variant using pairwise Jaccard's similarity indices. AMR gene variant pairs were defined as similar if they shared any 31-mer, resulting in a total of 46,279 possible similar AMR gene variant pairs (**Fig.S3-S5**).

For the *low coverage* scenario, reads were simulated from 176 bla_{TEM} gene-containing constructs at coverage depths ranging from 1x to 50x using ART (n=176*50=8,800 simulations), reflecting total bla_{TEM} diversity present in the ResFinder database at the time of simulation. Each construct contained a random perfect reference bla_{TEM} variant flanked by 1kb of random sequence on each side produced using Numpy/BioPython as above. Simulated reads were then processed by each genotyping method using default settings and the identified variants were compared with the known bla_{TEM} variants present in each construct. The measure of performance for this scenario was the proportion of bla_{TEM} variants correctly identified by each method at each coverage level.

Real data: Isolate selection

To evaluate performance on real data, we then studied a total of 1,818 *E. coli* isolates comprising three different WGS datasets in order to reflect different strain-level and AMR gene diversity: (i) 984 sequentially collected bloodstream infection isolates at Oxford University Hospitals (OUH) NHS Foundation Trust(19) ("Oxford dataset"); (ii) 497 animal commensal *E. coli* isolates donated by the UK Animal and Plant Health Agency (APHA)(20) ("APHA dataset"), and (iii) 337 *E. coli* isolates collected by UK Health Security Agency's (UKHSA) Antimicrobial Resistance and

Healthcare Associated Infections (AMRHAI) Reference Unit, which investigates isolates enriched for rare or important resistance genotypes encountered in the UK (sequenced for this study, "UKHSA dataset").

Isolates were re-cultured from frozen stocks stored in nutrient broth plus 10% glycerol at -80°C. DNA was extracted using the QuickGene DNA Tissue Kit S (Kurabo Industries, Japan) as per manufacturer's instructions, with an additional mechanical lysis step (FastPrep, MP Biomedicals, USA) immediately following chemical lysis. A combination of standard Illumina and in-house protocols were used to produce multiplexed paired-end libraries, which were sequenced on an Illumina HiSeq 2500, generating 151bp paired-end reads. High quality sequences were de-novo assembled using Velvet(21) as previously described(22). *In silico* Achtman(23) multi-locus sequence types (MLST) types were defined using ARIBA(6).

While this work does not attempt to predict resistance from WGS data, each isolate had linked AST (summarized in **Table S2, Fig.S6**), which we have included as the complexity of resistance genotype identification is associated with the phenotype. Isolates had complete AST data available for: ampicillin, ceftazidime and one other 3rd generation cephalosporin (cefotaxime for the animal commensal isolates, ceftriaxone for all others), gentamicin, ciprofloxacin, and co-trimoxazole.

We compared AMR genotypes reported for each isolate by each method, stratified by antibiotic class to which resistance was conferred as specified in the ResFinder database, namely: beta-lactams, aminoglycosides, quinolones, trimethoprim, and sulphonamides. Discrepancies were classified according to which of the four bioinformatics methods agreed (**Fig.S7**). The cause of discrepancy was investigated for all beta-lactam resistance genotypes, because these antibiotics are most commonly used for clinical *E. coli* infections, and then for discrepancy patterns occurring in >1.5% (n=27) of isolates for the other classes.

6. Results

Simulated scenarios

Accurate identification of single AMR gene variants in simulated sequence constructs

For the 3,092 AMR gene variants in the ResFinder database, all four genotyping methods correctly identified those inserted into random sequence contexts in 2,011 (63.5%) cases. KmerResistance was correct for all 3,092 simulations, ABRicate for 3,082 (99.7%), ARIBA for 2,927 (94.7%) and SRST2 for 2,120 (68.6%) (**Fig.2**). For SRST2, most errors were due to its approach of pre-clustering reference sequences into sub-families by sequence identity prior to genotyping, thereby essentially excluding *a priori* the possibility of identifying alleles that were not selected as the representative for these sub-family clusters. This error is explained in more detail below as it also affected genotyping in real isolate sequences.

Impact of the presence of multiple closely related alleles on genotyping calls

The multiple allele simulation caused significant problems for assembly-based algorithms, with ABRicate reporting fragmented/incomplete alleles for 32,194/46,279 (70%) simulations and ARIBA reporting no alleles meeting its assembly quality requirements for 32,987/46,279 (71%) simulations. SRST2, as expected, found only

a single allele in most (33077/46,279 (71%)) cases (**Table 1**), as dictated by its clustering parameters. ABRicate managed to identify both alleles correctly in the absence of incorrect calls in 11,382/46,279 (25%) of simulations, whereas ARIBA and SRST2 only managed to correctly reconstruct both members of the pair in the absence of correct calls in 2,494/46,279 (5%) and 2,539/46,279 (5%) cases respectively (Table 1). Of the four programs, KmerResistance performed the best, identifying both alleles correctly without additional erroneous calls in 38,826/46,279 (84%). of cases. Unsurprisingly all four programs were most likely to make erroneous genotyping calls as the simulated pairs of alleles became more closely related (**Fig.S8)**.

Impact of sequencing depth on genotyping calls

KmerResistance was able to identify *bla*$_{TEM}$ alleles at lower coverage than any of the other methods (**Fig.1**). Above 15x depth of coverage for the gene, all methods correctly identified *bla*$_{TEM}$ alleles in simulated constructs in > 95% of cases (**Fig.1**). All methods were able to identify all of the *bla*$_{TEM}$ alleles correctly at least once, but examples existed for all methods where the allele was correctly identified at low coverage, but then mis-classified at higher coverage. In general, ABRicate and SRST2, while requiring greater sequencing depth to correctly identify *bla*$_{TEM}$ alleles initially were more accurate at higher coverage depths, making erroneous calls for only 1/176 (0.6%) and 0/176 (0%) of *bla*$_{TEM}$ alleles at depths >20x. In contrast, for >20x coverage ARIBA and KmerResistance made erroneous allele calls for 23/176 (13%) and 6/176 (3%) *bla*$_{TEM}$ variants respectively. Above 40x coverage ABRicate was incorrect for one (0.6%), ARIBA for four (2%), KmerResistance for one (0.6%), and SRST2 for zero (0%) simulated *bla*$_{TEM}$ alleles.

Real data

E. coli isolate diversity, antimicrobial susceptibility phenotypes and antimicrobial resistance genotypes

The 1,818 isolates were diverse, representing >260 multi-locus sequence types (STs), which were differentially distributed among the datasets. For example, although ST131 was the most common (207/1818 (11%) isolates), this was largely due to the fact it was by far the most common in the UKHSA dataset (74/337 (22%) isolates). In the Oxford dataset, it was only the second most common ST (123/984 (13%) isolates) after ST73 (161/984 (16%)) isolates) and it was rare in the APHA isolates (10/497 isolates (2%)).

Correspondingly, the set also contained a broad range of resistance genes, but the exact number was dependant on the method of search. For legibility, we have included results as reported by ABRicate as this is the most conceptually simple and interrogatable approach.. The commonest AMR-associated sequence identified was *mdfA*. This is known to be universal in *E. coli*, and correspondingly was identified in all 1,818 isolates in the dataset. There were no other ubiquitous AMR genes; however, several were common across datasets, with bla$_{TEM}$, *aadA*, *sul*, *tet*, and *dfr* genes occurring in >40% of the isolates. As expected, more UKHSA isolates contained extended-spectrum beta-lactamase (54/337 vs 94/1481) and carbapenemase (18/337 vs 1/1481) genes (p=<0.001). Aside from bla$_{TEM}$, other beta-lactamases were rare among the APHA dataset. Outside of beta-lactam-

associated AMR genes, the Oxford dataset had the lowest proportion of other AMR genes for all the different gene families encountered in this study.

Genotyping discrepancies

10,487 different genes (N=15,588 different alleles) were identified in the 1818 isolates by the four methods. 1,392/1,818 (76%) isolates had discrepancies across the four bioinformatics methods for at least one gene. At the gene-level, aside from for *tet, aadA* and *cat* genes, the performance of the bioinformatic tools was similar (**Fig.3, panel a**), with tools reporting each gene in the approximately same proportion of isolates (within +/-2%). With regards to the three outliers, ABRicate reported *tet* and *aadA* genes in 19% and 10% more isolates respectively than the other three tools, and ABRicate and KmerResistance reported *cat* genes in 5% more isolates than ARIBA and SRST2. By contrast, the alleles reported by each tool were often discrepant, with alleles of some genes (e.g. *blaSHV*, *blaCMY*) consistently being differentially reported (**Fig.3, panel b**). Consequently, pairwise agreement between any two different tools was less than 59% (N=1,065 isolates, **Fig.3, panel c**). While unsupported genotype reports (i.e. where the output of one tool was not supported by any other) were common for all tools (**Fig.4**), KmerResistance reported fewer unsupported genotypes than the other three tools ($p<0.001$).

Causes of genotyping discrepancy

At least 2,530/4,321 (59%) of allele-level discrepancies were due to programs naming the same underlying sequence differently (annotation differences). We identified three major causes of differences through investigation of discrepantly reported genes: (i) difficulty distinguishing between optimal matches among alleles with nested sequences (N=1,737 genes); (ii) spurious identification of additional alleles due to reads being multiply mapped to distant variants of the same allelic family (N=547 genes); and (iii) tools choosing different optimal matches based on DNA sequence alignment when the database only contains one sequence per protein (N=197) (**Fig.5**). These issues occurred alone in 1,944/2,530 (77%) discrepantly reported genes, and or in combination in 586/2,530 (23%) cases. In isolation these errors typically caused only a single method to be discordant, but when combined resulted in more complex patterns of discrepancy and could make all four methods disagree with one another. In addition to annotation, ABRicate's more relaxed requirement for complete gene coverage (which aims to mitigate assembly errors) caused at least 1,430/4,321 (33%) allele-level discrepancies. Discrepancies less easily classified as (but likely related to) annotation/cut-offs did occur, but only affected 381/10487 (4%) of reported genotypes.

Annotation-related discrepancies

The most common type of annotation error (N=1,737 genes) was the result of tools struggling to choose optimal matches where the database contained nested sequences. One such example of this (N=24) was caused by the sequences for two different *dfrA7* alleles in the October 2019 Resfinder database, dfrA7_1_AB161450 and dfrA7_5_AJ419170. The shorter of the two (dfrA7_1_AB161450, 474 base pairs long) aligns almost perfectly (percentage identity = 99%, 1 single nucleotide gap) with the first 473 bases of dfrA7_5_AJ419170. ARIBA, KmerResistance and SRST2, which look for the best identity sequence matches, all report the sample contains a perfect match for dfrA7_1_AB161450. By contrast ABRicate, which uses BLAST to

identify optimal sequences, reports the sample contains a near perfect match to dfrA7_5_AJ419170, as with this being a longer match it is more statistically significant. Similar errors occurred for several other genes, including *sul*, *tet*, *aph(6)*, and *aac(3)*.

The second most common annotation discrepancy (N=547 genes) represented tools reporting multiple alleles due to reads mapping to two or more distant variants of the same allelic family. An example observed was ARIBA and SRST2 reporting multiple bla_{SHV} alleles. In this instance, ARIBA and SRST2 identified a primary perfect allele and a second allele with a lower quality match. These multiple matches however were likely spurious, with <10 reads mapping individually to each allele, no clear heterozygosity observed in read pileups, and no fragmentation in assembly graphs. This is the result of a biproduct of how mapping methods identify optimal matches. Both ARIBA and SRST2 map reads to each sequence in the database, and then compare “closely related” sequences to decide which mapping is optimal. Defining “closely related” however is not straightforward (**Fig.S9**). Reads mapping to more than one set of “closely related“ sequences can result in tools finding multiple gene variants when the isolate only had one gene original

The final common annotation discrepancy (N=197 genes) was due to allele reporting based on which sequence in the database had the optimal DNA alignment with the target resistance gene. Although resistance gene nomenclature is largely based on protein sequence, but resistance gene databases mostly only catalogue one nucleotide sequence linked to an associated protein sequence. Variant alleles with synonymous mutations fail to perfectly match any element, and may have an alternate optimal DNA match. We observed this on 9 occasions where ABRicate, KmerResistance and SRST2 identified imperfect nucleotide-level matches to aph(3")-Ib_2_AF024602 and ARIBA identified an imperfect match to aph(3")-Ib_4_AF313472. However, the sequence they were matching to in the SPAdes and ARIBA assembly was a 100% identity and coverage protein match to aph(3")-Ib_5_AF321551.

Non-annotation related discrepancies

In addition to annotation discrepancies that were caused by bioinformatics algorithms, genotyping calls were also affected by partial/low coverage of AMR gene targets and assembly fragmentation, consistent with the results from simulations. For some of these, such as the 1,430 cut-off related discrepancies occurring for *tet, mfs, aadA,* and *cat* genes, each program identified the same section of sequence, making it clear that the different programs had different thresholds for reporting, other situations were less clear. To investigate this in detail, we examined beta-lactamase matches which were either partial/low coverage or occurred across fragmented assemblies.

Partial/low coverage beta-lactamase genes were discrepantly found in 39 isolates (**Fig.S10**), particularly affecting bla_{TEM}-like gene calls (29/39 cases). KmerResistance reported the presence of a beta-lactamase gene in all 39 of these discrepant cases, with calls supported to a varying degree by the other algorithms. However, in all but four cases, KmerResistance reported that the depth of the gene was less than 5x. For the four cases where the gene was present at greater than 5x depth as called by

KmerResistance, three (present at depth >100x) were omitted from ARIBA reports as ARIBA assemblies contained mis-sense mutations and the final one (present at depth 17x) also failed to assemble for ABRicate.

Assembly fragmentation affected ABRicate and ARIBA beta-lactam resistance gene calls in 24 cases, with 16 of these likely to be due to the presence of multiple closely related beta-lactamase alleles affecting assembly integrity. The possibility of heterozygous alleles was indicated by the ARIBA flag "variants_suggest_collapsed_repeat", and the SRST2 "minor allele frequency value" was high (>20%). KmerResistance reported two related alleles in 12/16 cases, one with high depth, percentage identity and coverage, and one much less accurately. This likely reflects KmerResistance's winner-takes-all strategy, where matching unique k-mers to reference alleles are counted, and the reference allele with the most matches is then also assigned all reads with non-unique kmer-matches. This then leaves only reads with unique k-mers matching any closely related secondary allele, resulting in poor depth and coverage metrics.

7. Discussion

We evaluated the impact of bioinformatics approaches to AMR genotyping in *E. coli* for four commonly used methods and a widely used AMR gene database (ResFinder). Using >50,000 simulations and comparing >1,800 sequences sampled across human and animal reservoirs, thereby capturing common and rare AMR genotypes, we highlight that whilst currently available, widely-used genotyping methods are useful, their outputs should be carefully considered in light of our findings. Commonly postulated causes of discrepancy, such as low quality sequencing data, appeared to play little role. Instead, discrepancies were primarily artefactual, occurring because of different approaches in representing the complexity of the reference AMR gene database. Inconsistent labelling of gene variants will also affect the reliability of any catalogue-based methods for phenotypic prediction from WGS-based AMR genotypes. Specifically, predicting phenotype based on the presence of specific allelic variants will be problematic without a reliable method of identification.

- Our work agrees with previous findings by Doyle *et al.* on a small and selected dataset(13); however, we utilised large simulated and real-life datasets to identify these significant genotyping discrepancies between methods, and also characterized the underlying reasons for these discrepancies. We found most discrepancies were largely due to annotation differences, i.e. each method identified the same consensus sequence but then named them differently. Further, many of these discrepancies are caused by implicit and frequently incorrect assumptions about database structure and AMR gene diversity, namely: that AMR genes can be classified in well-defined families using genetic identity, that different approaches to deciding best-matching alleles are equivalent, and that isolates will usually not harbour highly genetically related variants of the same AMR gene. However, nomenclature and family structure amongst AMR genes relevant to Enterobacterales is complicated, with highly diverse genotypes (and sometimes phenotypes) being assigned similar family names (e.g. $bla_{\text{CTX-M}}$, bla_{OXA}) and single SNPs in some cases leading to different resistance phenotypes (e.g. $bla_{\text{TEM-1}}$ (Genbank: AY458016.1) -

beta-lactamase inhibitor susceptible i.e. susceptible to amoxicllin-clavulanate, *bla*$_{TEM-30}$ (Genbank: AJ437107.1) - beta-lactamase inhibitor resistant i.e. resistant to amoxicillin-clavulanate). Given this, it is not surprising that we found methods that make fewer assumptions (e.g. KmerResistance) to be more robust. Based on our findings accurate resistance genotyping may require the use of multiple different methods to cross-check results, and a clear understanding of the specific assumptions underlying the methods used, before conclusions about allele presence are drawn. The alternative is the development of new algorithms that cope better with underlying AMR gene diversity in these organisms.

One of the key strengths of this analysis was its combined use of both simulations and real world data. By using simulations, we were able to benchmark methods against a known truth, which is impossible to do with real-world data. Previous studies using only real-world data have attempted to overcome the absence of complete knowledge of the underlying genotype by using phenotypic data as a reference standard; however genotype-phenotype correlations remain poorly defined(10, 19). By subsequently using a large sequencing dataset of isolates obtained across niches, we were then able to assess the extent of discrepancies in real-life, replicating the problems observed in simulated data.

A limitation of this work is that we chose not to evaluate the impact of database choice, and this will represent future work. Currently, as has been highlighted previously(24), there are discrepancies between the AMR databases in common use, with each having a slightly different scope and in some cases differential names for different AMR gene variants (e.g.strA vs *aph(6)-Ia* or *aphD*, and *strB* versus *aph(6)-Id*). Comparing databases would have therefore added significant further complexity whilst limiting the generalisability of findings. A further limitation stemming from our fixed choice of database is that we have not analysed any methods where the bioinformatic method and database are intertwined (e.g. ResFinder/PointFinder or RGI). As the interaction between tool and database was the cause of many issues, it is possible that methods that are database-specific will perform better. However, the drawbacks of these combined resources are their inflexibility, again limiting generalisability. A further limitation was that these genotyping algorithms were compared using an older version of the ResFinder database – the most up to date when this work was originally planned. Since this time, 70 sequences have been added, 2 sequences modified and 2 sequences deleted (See supplementary data). We opted not to re-perform the analysis due to its manual nature and that as most of the discrepancies relate to underlying principles behind the algorithms rather than the specific implementation. Finally, we have focused our evaluation on *E. coli*, but it is likely that these issues will also more widely affect AMR genotyping, particularly of similar species with complex genotypes.

While WGS-based approaches are attractive for both characterizing AMR gene epidemiology and representing a subsequent tool for resistance prediction, this work highlights the need for caution when interpreting resistance genotypes reported by even widely used bioinformatics methods. Before WGS-based approaches can be considered reliable for use in *E. coli* (and likely other Enterobacterales), particularly for clinical decision making or replacing phenotypic data to determine

epidemiological trends, database standardisation, the development of novel genotyping approaches, and improved validation and evaluation will be required.

8. Author statements

8.1 Authors and contributors

TJD, NS, AES, ASW, DWC and TEAP conceptualised the study. TD, NS, ASW, AES and MFA decided the methodology. NS, ASW, MFA, AES, DWC and TEAP supervised the project. NS, MA, MFA, MJE, KH and SH acquired and curated the data used in this study. TJD and JSW constructed software pipelines to analyse sequencing data using each of the bioinformatic tools. TJD and ASW investigated the data. TJD performed the formal analysis. NS, AES, SL, HP, AES and TEAP assisted with interpreting the cause and impact of discrepancies. TJD and NS wrote the original draft. TJD, NS, AES, PWF, TEAP and ASW assisted with data visualisation. All authors were involved in the review and editing process.

8.2 Conflicts of interest

The authors have no conflicts of interest to declare.

8.3 Funding information

The study was funded by the National Institute for Health Research Health Protection Research Unit (NIHR HPRU) in Healthcare Associated Infections and Antimicrobial Resistance at Oxford University in partnership with Public Health England (PHE) [NIHR200915]. DWC, TEAP, PWF and ASW are supported by the NIHR Oxford Biomedical Research Centre. The report presents independent research funded by the National Institute for Health Research. The views expressed in this publication are those of the authors and not necessarily those of the NHS, the National Institute for Health Research, the Department of Health or Public Health England. NS is an Oxford Martin Fellow and an NIHR Oxford BRC Senior Fellow. ASW is an NIHR Senior Investigator.

8.4 Ethical approval

Not applicable.

8.5 Acknowledgements

We are grateful to the microbiology laboratory teams at the John Radcliffe Hospital, Oxford, the Animal and Plant Health Agency, and UK Health Security Agency.

9. References

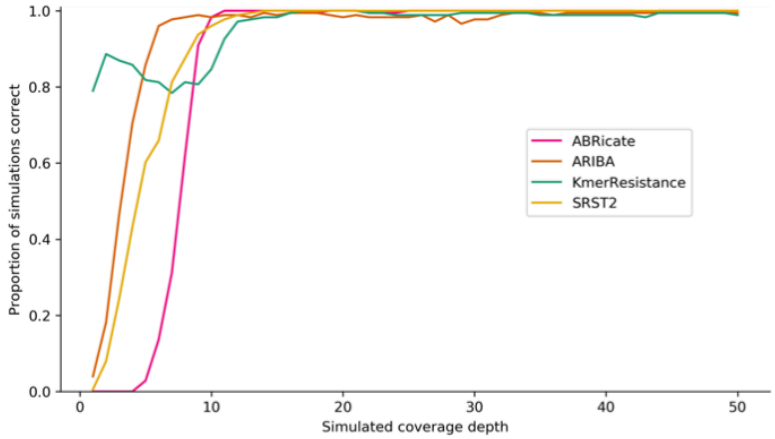
1. Quainoo S, Coolen JPM, van Hijum S, Huynen MA, Melchers WJG, van Schaik W, et al. Whole-Genome Sequencing of Bacterial Pathogens: the Future of Nosocomial Outbreak Analysis. Clin Microbiol Rev. 2017;30(4):1015-63.
2. Quan TP, Bawa Z, Foster D, Walker T, Del Ojo Elias C, Rathod P, et al. Evaluation of Whole-Genome Sequencing for Mycobacterial Species Identification and Drug Susceptibility Testing in a Clinical Setting: a Large-Scale Prospective Assessment of Performance against Line Probe Assays and Phenotyping. J Clin Microbiol. 2018;56(2).

3. Bortolaia V, Kaas RS, Ruppe E, Roberts MC, Schwarz S, Cattoir V, et al. ResFinder 4.0 for predictions of phenotypes from genotypes. J Antimicrob Chemother. 2020;75(12):3491-500.
4. Alcock BP, Raphenya AR, Lau TTY, Tsang KK, Bouchard M, Edalatmand A, et al. CARD 2020: antibiotic resistome surveillance with the comprehensive antibiotic resistance database. Nucleic Acids Res. 2020;48(D1):D517-d25.
5. Feldgarden M, Brover V, Haft DH, Prasad AB, Slotta DJ, Tolstoy I, et al. Validating the AMRFinder Tool and Resistance Gene Database by Using Antimicrobial Resistance Genotype-Phenotype Correlations in a Collection of Isolates. Antimicrob Agents Chemother. 2019;63(11).
6. Hunt M, Mather AE, Sánchez-Busó L, Page AJ, Parkhill J, Keane JA, et al. ARIBA: rapid antimicrobial resistance genotyping directly from sequencing reads. Microb Genom. 2017;3(10):e000131.
7. Inouye M, Dashnow H, Raven LA, Schultz MB, Pope BJ, Tomita T, et al. SRST2: Rapid genomic surveillance for public health and hospital microbiology labs. Genome medicine. 2014;6(11):90.
8. Clausen PT, Zankari E, Aarestrup FM, Lund O. Benchmarking of methods for identification of antimicrobial resistance genes in bacterial whole genome data. J Antimicrob Chemother. 2016;71(9):2484-8.
9. Zankari E, Allesøe R, Joensen KG, Cavaco LM, Lund O, Aarestrup FM. PointFinder: a novel web tool for WGS-based detection of antimicrobial resistance associated with chromosomal point mutations in bacterial pathogens. J Antimicrob Chemother. 2017;72(10):2764-8.
10. Stoesser N, Batty EM, Eyre DW, Morgan M, Wyllie DH, Del Ojo Elias C, et al. Predicting antimicrobial susceptibilities for Escherichia coli and Klebsiella pneumoniae isolates using whole genomic sequence data. J Antimicrob Chemother. 2013;68(10):2234-44.
11. Shelburne SA, Kim J, Munita JM, Sahasrabhojane P, Shields RK, Press EG, et al. Whole-Genome Sequencing Accurately Identifies Resistance to Extended-Spectrum β-Lactams for Major Gram-Negative Bacterial Pathogens. Clin Infect Dis. 2017;65(5):738-45.
12. Stubberfield E, AbuOun M, Sayers E, O'Connor HM, Card RM, Anjum MF. Use of whole genome sequencing of commensal Escherichia coli in pigs for antimicrobial resistance surveillance, United Kingdom, 2018. Euro Surveill. 2019;24(50).
13. Doyle RM, O'Sullivan DM, Aller SD, Bruchmann S, Clark T, Coello Pelegrin A, et al. Discordant bioinformatic predictions of antimicrobial resistance from whole-genome sequencing data of bacterial isolates: an inter-laboratory study. Microb Genom. 2020;6(2).
14. Seemann T. ABRicate. 2020.
15. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, et al. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J Comput Biol. 2012;19(5):455-77.
16. Harris CR, Millman KJ, van der Walt SJ, Gommers R, Virtanen P, Cournapeau D, et al. Array programming with NumPy. Nature. 2020;585(7825):357-62.
17. Cock PJ, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A, et al. Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics. 2009;25(11):1422-3.

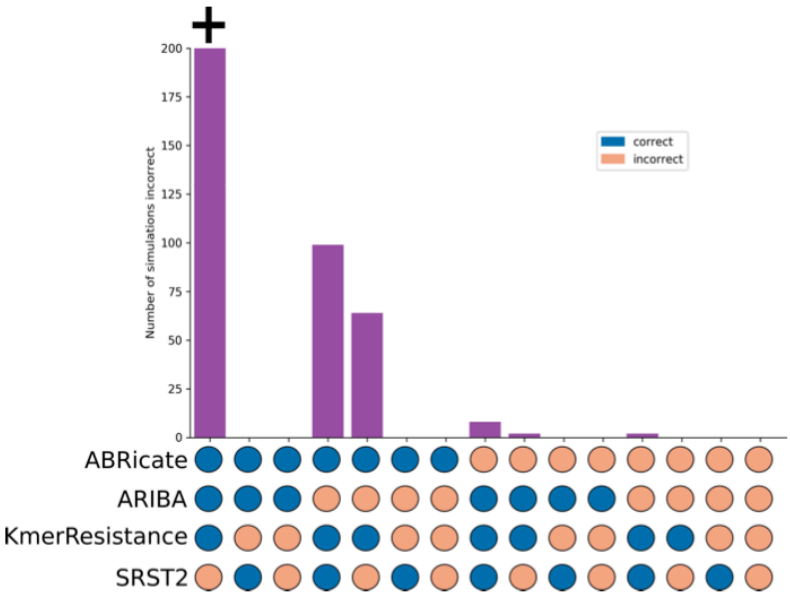
567 18. Ellington MJ, Ekelund O, Aarestrup FM, Canton R, Doumith M, Giske C, et al.
568 The role of whole genome sequencing in antimicrobial susceptibility testing of
569 bacteria: report from the EUCAST Subcommittee. Clin Microbiol Infect. 2017;23(1):2-
570 22.
571 19. Davies TJ, Stoesser N, Sheppard AE, Abuoun M, Fowler P, Swann J, et al.
572 Reconciling the Potentially Irreconcilable? Genotypic and Phenotypic Amoxicillin-
573 Clavulanate Resistance in Escherichia coli. Antimicrob Agents Chemother.
574 2020;64(6).
575 20. AbuOun M, O'Connor HM, Stubberfield EJ, Nunez-Garcia J, Sayers E, Crook
576 DW, et al. Characterizing Antimicrobial Resistant Escherichia coli and Associated
577 Risk Factors in a Cross-Sectional Study of Pig Farms in Great Britain. Front
578 Microbiol. 2020;11:861.
579 21. Zerbino DR. Using the Velvet de novo assembler for short-read sequencing
580 technologies. Curr Protoc Bioinformatics. 2010;Chapter 11:Unit 11 5.
581 22. Stoesser N, Sheppard AE, Peirano G, Anson LW, Pankhurst L, Sebra R, et al.
582 Genomic epidemiology of global Klebsiella pneumoniae carbapenemase (KPC)-
583 producing Escherichia coli. Sci Rep. 2017;7(1):5917.
584 23. Wirth T, Falush D, Lan R, Colles F, Mensa P, Wieler LH, et al. Sex and
585 virulence in Escherichia coli: an evolutionary perspective. Mol Microbiol.
586 2006;60(5):1136-51.
587 24. McArthur AG, Tsang KK. Antimicrobial resistance surveillance in the genomic
588 age. Ann N Y Acad Sci. 2017;1388(1):78-91.

10. Figures and tables

Figure 1. Proportion of correct genotype calls for single AMR gene variants in simulated constructs by coverage depth and bioinformatics method.



593 **Figure 2. Identification of known single AMR gene variants in simulated**
594 **contexts by bioinformatic method.** Note only cases where one or more methods
595 were incorrect are shown (n=1,081). "+" denotes the case where total SRST2-only
596 errors=906, but are truncated to 200 to make other errors visible. blue = method
597 correct for these simulations, orange = method incorrect.
598



599
600
601

Figure 3. Gene identification concordance vs allele identification concordance. a) The number of isolates containing at least one allele of the name gene families (x-axis) stratified by method. b) The proportion of times a given gene was identified concordantly by all four methods. c) Pairwise agreement between the different methods across all isolates.

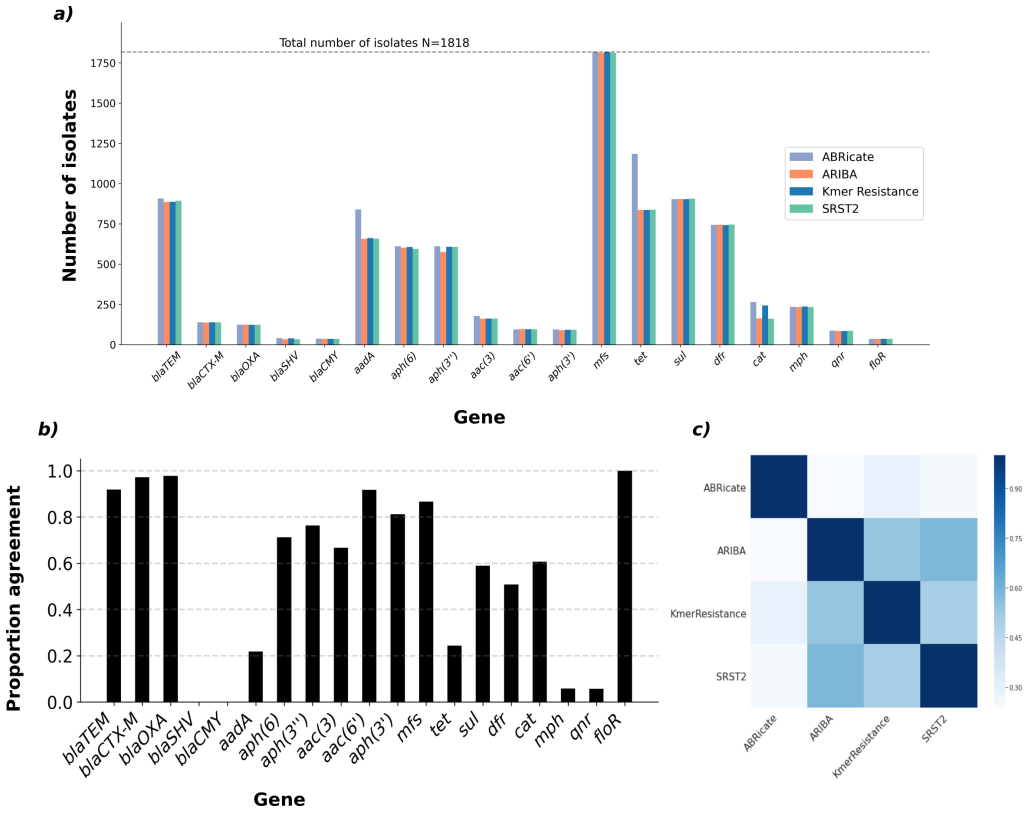


Figure 4. Genotype calls produced by a single method only, stratified by antibiotic class.

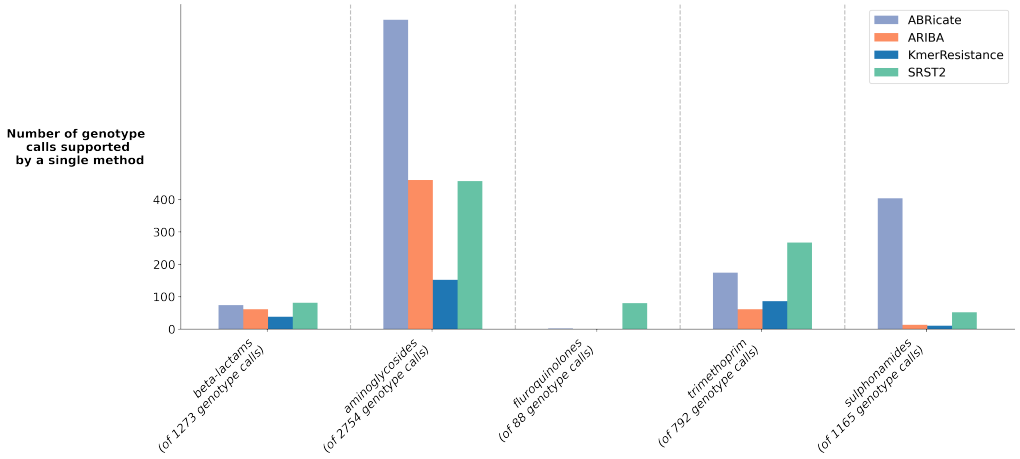
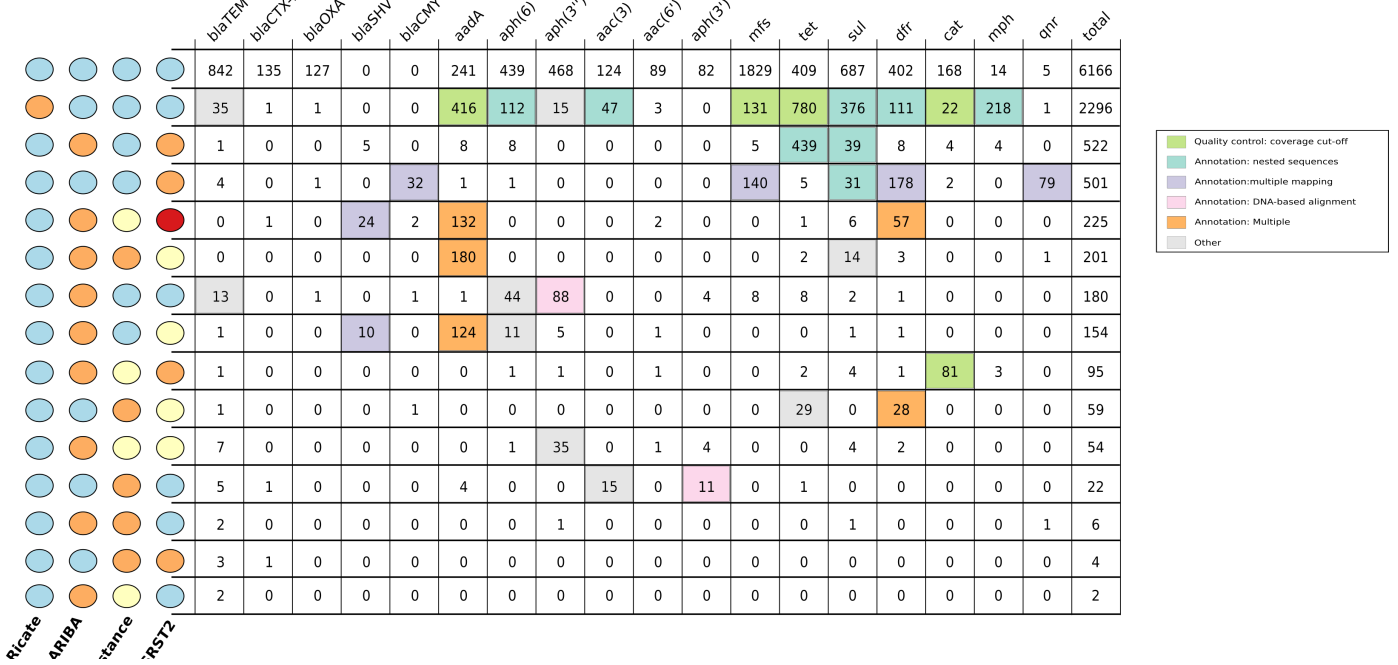


Figure 5. Genotyping agreement across all four bioinformatics algorithms, stratified by gene.
Colours on the left indicate which methods agreed with one another, with circles with the same colour indicating agreement. Colours in the main panel of the figure were used to identify the cause of the discrepancy, as denoted in the figure key. Cells (in the figure) were coloured if > 90% of isolates were caused by a given discrepancy. Cells with <10 isolates were not investigated.



624 **Table 1. Performance of genotyping methods in evaluating simulated**
625 **constructs with two related allelic variants.** Percentage reported out of a total of
626 46,279 simulations performed for each method.
627

	Number of calls (%)			
Genotyping call	**ABRicate**	**ARIBA**	**KmerResistance**	**SRST2**
No correct calls	17,145 (37%)	36,150 (78%)	489 (1%)	9,898 (21%)
One correct call but additional incorrect calls	2,419 (5%)	2 (0%)	1,452 (3%)	152 (0%)
One correct call, no incorrect calls	15,333 (33%)	7,634 (17%)	2,203 (5%)	33,077 (71%)
Two correct calls, but additional incorrect calls	0 (0%)	1 (0%)	3,309 (7%)	613 (1%)
Two correct calls, no incorrect calls	11,382 (25%)	2494 (5%)	33826 (84%)	2539 (5%)

628